**Editorial to accompany:**

Chiu Y, et al. Logistic early warning scores to predict death, cardiac arrest or unplanned intensive care unit re-admission after cardiac surgery. Anaesthesia 2019. doi:10.1111/anae.14755

**EDITORIAL**

# Improving early warning scores- more data, better validation, the same response

**K. J. Oglesby**[1], **J.A.C. Sterne**[2] and **B. Gibbison**[3]

*1 Specialty Registrar, Department of Intensive Care Medicine and Anaesthesia, University Hospitals Bristol NHS Foundation Trust, Bristol, UK*

*2 Professor, Department of Population Health Sciences, 3 Consultant Senior Lecturer, Bristol Heart Institute, University of Bristol, Bristol, UK*

**Correspondence to:** B. Gibbison

**Email:** ben.gibbison@bristol.ac.uk

**Keywords:** early warning scores; logistic regression; standardisation; cardiac surgery; ICU admission

**Twitter:** @kieranoglesby, @jonathansterne, @bengibbison

The practice of contemporary critical care is becoming synonymous with standardisation of clinical practice. This is apparent in the ever-expanding number of care bundles, consensus guidelines and protocolised care. Even strong advocates for physician autonomy or individualised patient management acknowledge the collective positive impact of these measures on both patient outcomes and healthcare efficiency [1]. Such standardisation has progressively spread beyond the walls of the intensive care unit (ICU), most obviously in the evolution of physiological observation 'track and trigger' systems, designed to detect and respond to critical illness on inpatient wards. Aggregated early warning scores (EWS) within these systems have evolved into the mutual language of ward and ICU teams. The fundamental principle of EWS is standardisation: assign numerical weightings to bedside observations; summate the overall degree of derangement; and mandate the timeframe for a clinical response.

In the UK, most institutions have chosen to implement the National Early Warning Score (NEWS) in preference to local EWS. At the launch of the latest iteration (NEWS-2) the review group noted that "*uptake of the NEWS across the NHS has exceeded all expectations*" [2]. This has been partially driven by support from influential bodies, including the National Institute for Health and Care Excellence (NICE) and NHS England. However, its extensive adoption by non-UK centres hints at the deeper strengths of NEWS. It is a characteristic example of standardisation in healthcare: a tool that is easy to use; decreases variation; increases efficiency; and facilitates communication. It performs at least as well as alternative EWS, within most clinical settings, most of the time [2]. This is excellent for a score based on consensus opinion rather than statistical models. The National Early Warning Score appears to be relatively good at predicting short-term mortality, normally classified as in-hospital death within 24 h [3,4], although attention has recently focussed on the performance of NEWS in predicting other critical illness outcomes. Events such as ICU re-admission, cardiac arrest, non-fatal organ failure and ICU resource utilisation appear to be less accurately predicted than 24-h in-hospital mortality [3]. These outcomes are important, as they frequently represent clinical situations where early intervention may have greater benefit.

**Commented [A1]:** This statement should have some supporting evidence referenced.

*Validation*

To date, validation of most EWS tools has been suboptimal. Validation of the NEWS quoted within the Royal College of Physicians' NEWS2 report [2] is based on a single centre on the south coast of England [4]. The demographic tested by this validation may not be applicable to other hospitals, particularly those in urban areas. Although much of the perceived benefit of EWS comes from a systematised approach, rigorous validation of these scoring systems prior to widespread use is

essential to optimise the benefit for patients and reduce the workload for staff. Most validation of NEWS has been in acute admissions and medical in-patients, settings in which it appears to perform well; there is less validity in elective surgical settings and non-medical specialities. These findings have encouraged researchers to develop specialty-specific scores and, within these subpopulations, the predictive value of a generic EWS suffers in comparison [5, 6]. The potential inference is that one standardised EWS may not apply with accuracy to all types of patients and all relevant critical illness events.

Validation of EWS models have generally focussed on discrimination (the ability to differentiate between patients who will and will not suffer an adverse event) rather than calibration (the degree of agreement between model predictions and the actual outcomes) [3, 7]. Calibration is most important when models are used in different populations and is often not reported for EWS, despite their use in varying population groups. Without confirmation of acceptable discrimination and calibration, the predictive accuracy of a model cannot be stated firmly [8]. Early warning scores developed using statistical analysis of routinely collected data representative of clinical practice appear to provide both improved discrimination and calibration, with potential for better and more efficient care than is possible [9] using consensus-based scores, such as NEWS-2.

The linked study by Chiu at al. in this issue of *Anaesthesia* may represent the next step towards future EWS, namely population-specific predictions developed using statistical analyses of large populations derived from electronic health records (EHR) [10]. Cardiac surgery in the UK has many good risk prediction tools, but these all aim to predict mortality pre-operatively (EuroSCORE-2 [11]) or in the first 24 h of ICU admission (ICNARC ARCtIC [12]). This new study analysed data from 13,631 patients who were discharged to the ward from ICU, extracted from an electronic vital signs charting system from four UK cardiac surgery centres. The centres varied in case-mix and workload. Logistic regression was used to model the relationship between bedside observations and subsequent adverse event (death, cardiac arrest or ICU admission) within 24 h. The performance of the resulting prediction score was compared with that of NEWS scores. Model validation on a subset of the dataset held out from model derivation ('internal validation') suggested that the logistic EWS achieved greater discrimination and better calibration than the NEWS. However, the improvement was only modest: for a score threshold of 5 during the previous 6 h, sensitivity increased from 48% to 52%, with no change in the specificity of 92%.

Importantly, Chiu et al. did not validate their score using external validation in a dataset separate from that used to develop the score. External validation of a model is important, because we expect worse performance in a new dataset than in a selected (hence representative) subset of the

development dataset. Given that NEWS was not derived from the dataset analysed by Chiu et al., the performance of their EWS is expected to be somewhat better. It remains possible that the performance of NEWS and the logistic EWS would be similar in an external dataset.

The event rates in the dataset analysed by Chiu et al. were low and re-admission to ICU was the most frequent event within the composite outcome and therefore dominates the model predictions. Re-admission to ICU is sensitive to the varying cultures, behaviours and resources between different hospitals using EWS. For example, one in 25 patients in Papworth are re-admitted to ICU, compared with 1 in 45 in Wolverhampton. Differences in case-mix will not fully account for this variation. Further, the logistic EWS predicts only subsequent re-admission to ICU, rather than all clinically meaningful deteriorations requiring intervention.

**Commented [A2]:** Needs reference.

**Commented [A3R2]:** This is data from the Chiu et al paper from which the editorial is about.

*Big data*

Using EHRs to collect large volumes of data from multiple sites represents the future in the development of EWS. As more UK hospitals adopt EHRs and standardise data definitions, the sample size and generalisability of EWS based on very large, combined datasets will increase. Linking physiological data to robust clinical outcomes stored elsewhere in the EHR is likely to improve the granularity and accuracy of predictions. To do this on a large scale requires national health and care data repositories. An early example of this is the National Institute of Health Research Critical Care Health Informatics Collaborative [13], which combines ICU data from seven UK centres (including physiological observations, drugs, interventions and outcomes at hourly intervals) that can be linked to Office for National Statistics and NHS Digital records. Were large numbers of UK hospitals to be included, then the opportunities for healthcare research would be myriad.

**Commented [A4]:** NHS Digital is the name of a body that deals with digital records of the NHS and therefore needs a capital letter.

The NHS has advocated a 'bottom-up' approach to information technology solutions – that NHS trusts should choose their own EHR systems. To reap the potential research benefit of EHRs, these systems must communicate at the 'back end', even if the 'front end' between different software suppliers looks different. Ideally, any repository would use a federated model, where centres classify and store their EHR data in a standardised way, with access to a central 'port' allowing retrieval of the required data from local sites.

Machine-learning techniques may improve predictions when based on very large, detailed clinical databases. Applications of machine-learning are beginning to appear [14]. A sample size of 13,000 patients, as in the data analysed by Chiu et al., is at the lower end of when machine-learning is useful compared with standard regression modelling. Machine-learning avoids strong parametric

assumptions, though this may be at the expense of being able to succinctly present or explain the patterns behind underlying predictions [15]. It is vital that predictions based on machine-learning are validated before they are implemented in clinical practice.

Chiu et al. should be applauded for developing an EWS driven by data, rather than using an expert-opinion based approach. However, before logistic EWS models can replace current, consensus-based scores such as NEWS2, clinically important performance improvements, based on robust external validation must be demonstrated. This will require large datasets, representative of the clinical setting within which the EWS will be used, which in turn requires the infrastructure to store and retrieve relevant data. This will only be possible with investment and a collaborative approach at a regional and national scale from healthcare providers, academic institutions and industry. It is critical that with the development of increasingly sophisticated EWS models, we retain an end-user experience that is simple and standardised. Accurate model predictions are of no use with an inadequate response from the clinical team. To this end, widespread dissemination, training and behavioural interventions are vital to realise the potential benefits of future specialty-specific EWS.

**Acknowledgements**

BG and JS are supported by the UK NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the UK Department of Health. No other conflicts of interest declared.

**References**


1 Hasibeder WR. Does standardization of critical care work? *Current Opinion in Critical Care* 2010; **16**: 493-8.

2 Royal College of Physicians. *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*. Updated report of a working party. London: RCP, 2017.

3 Smith ME, Chiovaro JC, O'Neil M, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Annals of the American Thoracic Society* 2014; **11**: 1454-65.

4 Smith GB, Prytherch DR, Meredith P, et al. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; **84**: 465-70.

5 Downey CL, Tahir W, Randell R et al. Strengths and limitations of early warning scores: A systematic review and narrative synthesis. *International Journal of Nursing Studies* 2017; **76**: 106-119.

6 Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical patients in the postoperative setting: retrospective cohort study. *Annals of Surgery* 2019; **269**: 1059-63.

7 Gerry S, Birks J, Bonnici T, et al. Early warning scores for detecting deterioration in adult hospital patients: a systematic review protocol. *BMJ Open* 2017; **7**: e019268.

8 Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine* 2015; **162**: 55-63.

9 Linnen DT, Escobar GJ, Hu X, et al. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. *Journal of Hospital Medicine* 2019; **14**: 161-69.

10. Chiu Y, Villar SS, Brand JW, Patteril MV, Morrice DJ, Clayton J, Mackay JH. Logistic early warning scores to predict death, cardiac arrest or unplanned intensive care unit re-admission after cardiac surgery. *Anaesthesia* 2019. doi:10.1111/anae.14755

11 Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. *European Journal of Cardiothoracic Surgery* 2012; **41**: 734-44.

12 Shahin J, Ferrando-Vivas P, Power GS, et al. The Assessment of Risk in Cardiothoracic Intensive Care (ARCtIC): prediction of hospital mortality after admission to cardiothoracic critical care. *Anaesthesia* 2016; **71**: 1410-16.

13 National Institute for Health Research. Critical Care - Health Informatics Collaborative, 2019. https://hic.nihr.ac.uk/?page_id=55 (accessed 25/07/2019).

14 Kwon JM, Lee Y, Lee Y, et al. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. *Journal of the American Heart Association* 2018; **7**: e008678.

15 Fralick M, Colak E, Mamdani M. Machine Learning in Medicine. *New England Journal of Medicine* 2019; **380**: 2588-89.